CMPUT 466/566: Machine Learning, Winter 2024
Tutorial 2

Shuai Liu

1 Likelihood function

We are in the setting of statistical inference which is also called learning in computer science. It studies the question of, given samples from a distribution:

$$X_1, ..., X_n \sim F,$$

how do we infer F? Here $X_1, ..., X_n$ are your data and F is your true underlying distribution. Up to this point in the course, we are under the setting of **parametric inference**, that is, we are assuming the pdf (or pmf) of the data distribution takes the form of $f(x;\theta)$. Formally, we assume that there exists a parameter set Θ such that

$$F(x) \in \{f(x;\theta) : \theta \in \Theta\}$$

Example 1. *If we assume F is a Gaussian distribution, then $\Theta = \mathbb{R} \times \mathbb{R}^+$ and $F \in \mathcal{P} := \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\} = \{\mathcal{N}(\mu, \sigma^2) : (\mu, \sigma) \in \Theta\}$.*

You may notice that there are two "inputs" to the function f: x and θ. θ is the **parameter** which uniquely identifies a single distribution if we restrict that all distributions under consideration takes the form of $f(x;\theta)$. In the above example, given a mean $\mu_1 \in \mathbb{R}$ and a standard deviation $\sigma_1 \in \mathbb{R}^+$, we uniquely identify a Gaussian distribution with the tuple (μ_1, σ_1). Mathematically, there exists a bijection between $\Theta = \mathbb{R} \times \mathbb{R}^+$ and $\mathcal{P}$.

Once we identify the distribution with the parameters, for each value that x can take, $f(x;\theta)$ assigns a probability density / probability mass. So here x is the argument, or **input**, of $f(\cdot;\theta)$. In order to distinguish between the input and the parameter, in the notation, we usually separate them with a semi-colon.

Back to the parametric inference setting. The parameters are known to the true underlying distribution, but not known to us, computer scientists. We would like to infer (or estimate) the true parameters, which uniquely identifies a distribution, given samples from this distribution. A common method is Maximum Likelihood Estimation (MLE), as you have seen in the class. Now we introduce the definition of likelihood function.

Definition 1. *Let $X_1, ..., X_n$ be i.i.d. with pdf/pmf $f(x;\theta)$ and x_i is the sample of X_i for all $i = 1, 2, ..., n$. The likelihood function is defined by*

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(x_i;\theta),$$

and the log-likelihood is defined by

$$\ell_n(\theta) = \mathcal{L}_n(\theta).$$

Remark 1. Parameter v.s. Input: for pdf/pmf, $f(x;\theta)$, it is a function of x. Eg. In the standard normal distribution $\mathcal{N}(0, 1)$, we fix mean to be 0 and variance to be 1. Then as we vary $x \in \mathbb{R}$, a probability density $f(x) = \frac{1}{\sqrt{2\pi}}(-\frac{x^2}{2})$ is assigned. In this case, the parameter is $(\mu, \sigma) = (0, 1)$ and the input is x.

Remark 2. The likelihood function is nothing but joint pdf/pmf of $X_1, ..., X_n$. However, it is viewed as a function of θ!

Example 2 (MLE of Bernoulli distribution)**.** *Let $X_1, ..., X_n$ be i.i.d. from* Ber(p)*, i.e., $f(x;p) = p^x(1-p)^{(1-x)}$. Let $S = \sum_{i=1}^n x_i$. Note that $S \leq n$. Then the likelihood function*

$$\begin{aligned}\mathcal{L}_n(p) &= \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} \\ &= p^{\sum_{i=1}^n x_i}(1-p)^{\sum_{i=1}^n (1-x_i)} \\ &= p^S(1-p)^{n-S}\ell_n(\theta) = S\log p + (n-S)\log(1-p).\end{aligned}$$

With some calculation, we can see that $\ell_n''(p) = -(\frac{S}{p^2} + \frac{(n-S)}{(1-p)^2}) \leq 0$, which implies that $\ell_n(p)$ is concave. Taking the derivative and setting it to be 0 we have that

$$\begin{aligned}\frac{d}{dp}\ell_n(\theta)\Big|_{p=\hat{p}} &= 0 \\ \frac{S}{\hat{p}} - \frac{n-S}{1-\hat{p}} &= 0 \\ \hat{p} &= \frac{S}{n}\end{aligned}$$

2 Random Vectors and Covariance Matrices

Definition 2. *A vector $\mathbf{X} = (X_1, ..., X_n)^\top$ whose components are scalar-valued random variables is called a random vector, or a multivariate random variable, and its distribution is called a multivariate distribution. The expectation of $\mathbf{X}$ is $\mathbb{E}[\mathbf{X}] = (\mathbb{E}[X_1], ..., \mathbb{E}[X_n])^\top \in \mathbb{R}^d$.*

Definition 3 (Covariance between random variables)**.** *The covariance between two random variables X, Y is defined to be*

$$\text{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

*(**exercise**: prove the last equality).*

Remark 3. Covariance intuitively reflects how Y changes when X is increased: if Y increases, then the covariance is positive, otherwise negative. Note that by definition $\text{Cov}(X,X) = \text{Var}(X)$ and $\text{Cov}(X,Y) = \text{Cov}(Y,X)$ for all random variables X, Y.

Proposition 1. *For two random variables X, Y (not necessarily independent), it follows that*

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y).$$

Proof. Exercise. □

Proposition 2. *The covariance matrix operator* Cov$(\cdot,\cdot)$ *is bilinear, that is, for random variable A, B, C,*

$$\text{Cov}(A, B+C) = \text{Cov}(A,B) + \text{Cov}(A,C).$$

Remark 4. If X is independent from Y, then $\text{Cov}(X,Y) = 0$ but the converse is not true. Consider X, Y such that

$$\mathbb{P}(X=-1) = \mathbb{P}(X=1) = 1/2, Y = X^2.$$

Here is a simple intuition about covariance matrix. Recall that variance characterizes how dispersed a distribution is. When we move from univariate random variable to multivariate random vectors, the covariance between different components of the random vector comes to play. Let $\mathbf{X} = (X_1, ..., X_n)^\top$. The dispersion of $\mathbf{X}$ can be affected by all of them and we understand how they affect the dispersion pairwisely by covariances.

Definition 4 (Covariance matrix)**.** *The covariance matrix is defined to be* $\mathbf{\Sigma} = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top]$ *or equivalently,* $\mathbf{\Sigma}_{ij} = \mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i) = \mathbf{\Sigma}_{ji}$*. Hence* $\mathbf{\Sigma}$ *is a symmetric matrix*

Proposition 3. *For all random vector* $\mathbf{X} \in \mathbb{R}^d$*, the covariance matrix* $\mathbf{\Sigma}$ *of* $\mathbf{X}$ *is a positive semi-definite matrix, that is, for all* $a \in \mathbb{R}^d$*,* $a^\top \mathbf{\Sigma} a \geq 0$*.*

Proof. For all $a \in \mathbb{R}^d$, we have that

$$\begin{aligned} a^\top \mathbf{\Sigma} a &= a^\top \mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top\right] a \\ &= \mathbb{E}\left[a^\top (\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top a\right] && \text{(Linearity of expectation)} \\ &= \mathbb{E}\left[((\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top a)^2\right] \geq 0 \end{aligned}$$

□

Corollary 1. *For a* 0*-mean random vector* $\mathbf{X} \in \mathbb{R}^d$*, its covariance matrix* $\mathbf{\Sigma}$ *is not positive-definite if and only if there exists a non-zero vector* $a \in \mathbb{R}^d$ *such that* $a^\top \mathbf{X} = 0$*, that is, there exists an index* $k \in \{1, 2, ..., d\}$ *such that* $X_k = \sum_{i \neq k} \frac{a_i}{a_k} X_i$*. In this case, we call* $X_1, ..., X_k$ *are linearly dependent.*

Proof. Since $\mathbb{E}[\mathbf{X}] = 0$, the covariance matrix $\mathbf{\Sigma}$ equals to $\mathbb{E}[\mathbf{X}\mathbf{X}^\top]$. By Proposition 3, $\mathbf{\Sigma}$ is PSD. It is not positive-definite if and only if there exists a non-zero vector $a \in \mathbb{R}^d$ such that $a^\top \mathbf{\Sigma} a = 0$, that is,

$$\begin{aligned} \mathbf{\Sigma} \text{ not positive-definite} &\iff \exists a \neq 0 \in \mathbb{R}^d, \text{ such that } a^\top \Sigma a = 0 \\ &\iff a^\top \mathbb{E}[X X^\top] a = 0 \\ &\iff \mathbb{E}[(a^\top X)^2] = 0 \\ &\iff a^\top X = 0 \end{aligned}$$

The second part of the statement follows by noting that $a \neq 0$. □

Remark 5. We sometimes call a general real random vector $\mathbf{Y}$ to be degenerated if the random variables $Y_1 - \mathbb{E}[Y_1], ..., Y_k - \mathbb{E}[Y_k]$ are linearly dependent.

Corollary 2. *For a random vector* $\mathbf{X} \in \mathbb{R}^d$ *and its covariance matrix* $\mathbf{\Sigma}$*, then* $\det(\mathbf{\Sigma}) > 0$ *if and only if* Y *is not degenerated.*

Proof. Exercise. □

Proposition 4. *For a random vector* $\mathbf{Z} \in \mathbb{R}^d$ *with mean* μ*, covariance matrix* $\mathbf{\Sigma}$ *and a matrix* $\mathbf{A} \in \mathbb{R}^{k \times d}$*, the vector* $\mathbf{AZ}$ *has mean* $\mathbf{A}\mu$ *and covariance matrix* $\mathbf{A\Sigma A}^\top$

Proof.

$$\begin{aligned} \mu_{\mathbf{AZ}} &= \mathbb{E}[\mathbf{AZ}] = \mathbf{A}\mathbb{E}[\mathbf{Z}] = \mathbf{A}\mu \\ \mathbf{\Sigma}_{\mathbf{AZ}} &= \mathbb{E}[(\mathbf{AZ} - \mathbb{E}[\mathbf{AZ}])(\mathbf{AZ} - \mathbb{E}[\mathbf{AZ}])^\top] \\ &= \mathbb{E}[\mathbf{A}(\mathbf{Z} - \mathbb{E}[\mathbf{Z}])(\mathbf{Z} - \mathbb{E}[\mathbf{Z}])^\top \mathbf{A}^\top] \\ &= \mathbf{A}\mathbb{E}[(\mathbf{Z} - \mathbb{E}[\mathbf{Z}])(\mathbf{Z} - \mathbb{E}[\mathbf{Z}])^\top]\mathbf{A}^\top \\ &= \mathbf{A\Sigma A}^\top \end{aligned}$$

□

3 Univariate Gaussian

There are lots of nice properties of Gaussians. We first show two useful properties for independent Gaussians.

Proposition 5. *Let $X \sim \mathcal{N}(\mu, \sigma^2)$ be a Gaussian random variable, then $X + c \sim \mathcal{N}(\mu + c, \sigma^2)$.*

Proof. Let $Y = X + c$ and $\Phi_X(x) = \mathbb{P}(X \leq x)$, $f_X(x)$ be the cdf and pdf of X respectively. Then Φ_Y the cumulative distribution function of Y is

$$\begin{aligned}\Phi_Y(y) = \mathbb{P}(Y \leq y) &= \mathbb{P}(X + c \leq y) \\ &= \mathbb{P}(X \leq y - c) \\ &= \Phi_X(y - c)\end{aligned}$$

Recall $f_Y(\cdot)$, the pdf of Y, is the derivative of $\Phi_Y(\cdot)$, that is,

$$\begin{aligned}f_Y(y) = \frac{d}{dy}\Phi_Y(y) &= \frac{d}{dy}\Phi_X(y - c) \\ &= f_X(y - c) \\ &= \frac{1}{\sigma\sqrt{2\pi}} \exp(-(y - c - \mu)^2/(2\sigma^2))\end{aligned}$$

□

Proposition 6. *Let $X \sim \mathcal{N}(\mu, \sigma^2)$ be a Gaussian random variable, thencX is a Gaussian with mean $\mathcal{N}(c\mu, c^2\sigma^2)$ for $c \neq 0$.*

Proof. Similar to the last proof. Exercise. □

Proposition 7. *Let $X \sim \mathcal{N}(\mu_1, \sigma_1^2), Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ be independent, then $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.*

Proof. For simplicity of calculation, we show the above proposition with $\mu_1 = \mu_2 = 0, \sigma_1 = \sigma_2 = 1$. The general case is left as an exercise. As before, we denote $\Phi_X(x)$, $\Phi_Y(y)$, $f_X(x)$, $f_Y(y)$ as the cdfs and pdfs of X and Y respectively. The following trick is often referred as the convolution trick. Let $Z = X + Y$ and the cdf $\Phi_Z(z)$ of Z can be written as

$$\begin{aligned}\mathbb{P}(Z \leq z) = \mathbb{P}(X + Y \leq z) &= \mathbb{P}(Y \leq z - X) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f_{XY}(x, y) dy dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f_X(x) f_Y(y) dy dx \\ &= \int_{-\infty}^{\infty} f_X(x) \left(\int_{-\infty}^{z-x} f_Y(y) dy \right) dx \\ &= \int_{-\infty}^{\infty} f_X(x) \Phi_Y(z - x) dx\end{aligned}$$

Taking the derivative of $\Phi(z)$ and using Lebniz's rule to interchange the order of derivative and integral, we obtain that

$$\begin{aligned}f_Z(z) = \frac{d}{dz}\mathbb{P}(Z \leq z) &= \frac{\partial}{\partial z} \int_{-\infty}^{\infty} f_X(x) \Phi_Y(z - x) dx \\ &= \int_{-\infty}^{\infty} f_X(x) \left(\frac{\partial}{\partial z} \Phi_Y(z - x) \right) dx \\ &= \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) dx\end{aligned}$$

Plug in the expression of $f_X(x), f_Y(y)$,

$$\begin{aligned}
f_Z(z) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2}\right)\cdot\exp\left(-\frac{(z-x)^2}{2}\right)dx \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left(-\frac{x^2+(z-x)^2}{2}\right) \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left(-\frac{2x^2-2zx+z^2}{2}\right)dx \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left(-\frac{2x^2-2zx+z^2}{2}\right)dx \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left(-x^2-zx-\frac{1}{4}z^2-\frac{1}{4}z^2\right)dx \\
&= \frac{1}{2\pi}\exp\left(-\frac{1}{4}z^2\right)\int_{-\infty}^{\infty} -x^2-zx-\frac{1}{4}z^2dx \\
&= \frac{1}{2\pi}\exp\left(-\frac{1}{4}z^2\right)\int_{-\infty}^{\infty} -\left(x-\frac{1}{2}z\right)^2dx \\
&= \frac{1}{2\pi}\exp\left(-\frac{1}{4}z^2\right)\cdot\sqrt{\frac{1}{2}}\cdot\sqrt{2\pi}
\end{aligned}$$

where the last equality follows by the pdf of $\mathcal{N}(0.5z, 0.5)$ integrates to 1, that is, $\int_{\mathbb{R}} \frac{1}{\sqrt{1/2}\sqrt{2\pi}}\exp(-(x-0.5z)^2)dx = 1$. □

Exercise. Prove Proposition 7 for the general case.

Corollary 3. *Let $X_1, ..., X_n$ be n independent Gaussian with mean $\mu_1, ...\mu_n$ and variance $\sigma_1^2, ..., \sigma_n^2$. Then $\sum_{i=1}^n c_iX_i \sim \mathcal{N}(\sum_{i=1}^n c_i\mu_i, \sum_{i=1}^n c_i^2\sigma_i^2)$ for constants $c_1, ..., c_n$ such that $\prod_{i=1}^n c_i \neq 0$.*

Proof. Without loss of generality, we assume $c_1, ..., c_n \neq 0$. Then for each i, c_iX_i is a Gaussian $\mathcal{N}(c_i\mu_i, c_i^2\mu_i^2)$ by Proposition 6. Applying Proposition 7 for $n-1$ times we obtain the stated result. □

We now move to the multivariate case. A random vector is defined to be multivariate Gaussian if it can be written as an affine transformation of a random vector with each component being i.i.d. standard Gaussian. We formalize this definition below.

4 Multivariate Gaussian

Definition 5 (Multivariate Gaussian)**.** *Let $U_i \sim \mathcal{N}(0,1)$ be i.i.d. standard Gaussian random variables. A random vector $\mathbf{Z} \in \mathbb{R}^k$ is a jointly Gaussian (JG) random vector with distribution being multivariate Gaussian (MG), if there exists a matrix $\mathbf{R} \in \mathbb{R}^{k\times l}$ and a vector $\mu \in \mathbb{R}^l$ such that $\mathbf{Z} = \mathbf{R}\mathbf{U} + \mu$ where $\mathbf{U} = (U_1, ..., U_l)^\top$.*

There are multiple equivalent definitions of MG that may be useful in different cases.

Theorem 1 (Equivalent definitions of JG)**.** *Let $\mathbf{Z} = (Z_1, ..., Z_k)^\top$ be a random vector. The followings are equivalent (TFAE):*

1. *$\mathbf{Z}$ is JG*
2. *For every non-zero $a = (a_1, ..., a_k)^\top \in \mathbb{R}^k$, the random variable $\sum_{i=1}^k a_iZ_i$ is Gaussian.*

3. (Non-degenerate case only) The pdf of $\mathbf{Z}$ *is*

$$f_{\mathbf{Z}}(\mathbf{z}) = \frac{1}{\sqrt{\det(\mathbf{\Sigma})}} \frac{1}{(\sqrt{2\pi})^k} \exp\left(-\frac{1}{2}(\mathbf{z}-\mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{z}-\mu)\right)$$

where $\mathbf{\Sigma} = \mathbb{E}[(\mathbf{Z}-\mu)(\mathbf{Z}-\mu)^\top] = \mathbf{E}[(\mathbf{RU})(\mathbf{RU})^\top] = \mathbf{R}\mathbb{E}[\mathbf{U}\mathbf{U}^\top]\mathbf{R}^\top = \mathbf{R}I\mathbf{R}^\top = \mathbf{R}\mathbf{R}^\top$

Exercise. Show that for all positive definite matrices M, there exists a distribution such that M is the covariance matrix of it.

Let's take a glance at some well-known properties of MG. Let $\mathbf{Z} \sim \mathcal{N}(\mu_{\mathbf{z}}, \mathbf{\Sigma}_{\mathbf{z}})$ be a JG random vector.

Exercise. Show that the linear transformation $\mathbf{AZ}$ is also JG where $\mathbf{A}$ has the dimensions scaled appropriately with the dimension of $\mathbf{Z}$.

$$\mathbf{AZ} \sim \mathcal{N}(\mathbf{A}\mu_{\mathbf{z}}, \mathbf{A}\mathbf{\Sigma}_{\mathbf{z}}A^\top)$$

Hint:(Use definition of JG and Proposition 4.)

Proposition 8. *Assume the partition of a JG* $\mathbf{Z} = [\mathbf{X}, \mathbf{Y}]^\top$ *whose distribution is given by* $\mathbf{Z} \sim \mathcal{N}(\mu_{\mathbf{z}}, \mathbf{\Sigma}_{\mathbf{z}})$ *and*

$$\mu_{\mathbf{z}} = [\mu_{\mathbf{X}}, \mu_{\mathbf{Y}}]^\top, \mathbf{\Sigma}_{\mathbf{Z}} = \begin{bmatrix} \mathbf{\Sigma}_{\mathbf{XX}} & \mathbf{\Sigma}_{\mathbf{XY}} \\ \mathbf{\Sigma}_{\mathbf{YX}} & \mathbf{\Sigma}_{\mathbf{YY}} \end{bmatrix}.$$

Then the conditional distribution of $\mathbf{X}$ *given* $\mathbf{Y}$ *is JG*

$$\mathbf{X}|\mathbf{Y} \sim \mathcal{N}(\mu_{\mathbf{X}} + \mathbf{\Sigma}_{\mathbf{XY}}\mathbf{\Sigma}_{\mathbf{YY}}^{-1}(\mathbf{Y}-\mu_{\mathbf{Y}}), \mathbf{\Sigma}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{YY}}^{-1}\mathbf{\Sigma}_{\mathbf{YX}})$$

and the marginal distribution of $\mathbf{X}$ *is also JG.*

Exercise. Show that the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ is JG and the marginal distribution of $\mathbf{Y}$ is also JG. *(Hint: Use definition of JG.)*

Remark 6. The converse is not necessarily true! If $\mathbf{X}, \mathbf{Y}$ are individually Gaussian, then we know that we can write $\mathbf{X}, \mathbf{Y}$ as linear transformations of i.i.d Gaussians $\mathbf{X} = \mathbf{R}_1\mathbf{U}_1$ and $\mathbf{Y} = \mathbf{R}_2\mathbf{U}_2$. But whether $[\mathbf{U}_1, \mathbf{U}_2]^\top$ are i.i.d random variables remain unclear.

Corollary 4. *If* $\mathbf{X}$ *and* $\mathbf{Y}$ *as described above are uncorrelated, that is* $\mathbf{\Sigma}_{\mathbf{XY}} = \mathbf{\Sigma}_{\mathbf{YX}} = \mathbf{0}$*, then they are independent.*

Proof. The conditional distribution of $\mathbf{X}|\mathbf{Y}$ can be written as

$$\mathbf{X}|\mathbf{Y} \sim \mathcal{N}(\mu_{\mathbf{X}} + 0 \cdot \mathbf{\Sigma}_{\mathbf{YY}}^{-1}(\mathbf{Y}-\mu_{\mathbf{Y}}), \mathbf{\Sigma}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{YY}}^{-1} \cdot 0) = \mathcal{N}(\mu_{\mathbf{X}}, \mathbf{\Sigma}_{\mathbf{XX}})$$

from the pdf of MG, we can get that

$$f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{Z}}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{Y}}(\mathbf{y})$$

(check it!). □

Remark 7. Recall that in general, uncorrelated does not imply independent!

Proposition 9. *Let* X, Y *be univariate Gaussian, then* $[X, Y]^\top$ *is JG if and only if the marginal distribution of* X*, the marginal distribution of* Y*,* $X|Y = y$ *and* $Y|X = x$ *are Gaussian.*

Example 3. *Let* X_1 *and* X_2 *be i.i.d. standard Gaussians. Let* U *be random variable uniformly distributed on* $\{-1, 1\}$*, independent of* X_1, X_2*. The we have* $[Z_1, Z_2]^\top$ *are JG for* $Z_1 = X_1, Z_2 = X_1 + X_2$ *using the above proposition. We verify that they are marginally Gaussian by Proposition 7:* $Z_1 \sim \mathcal{N}(0, 1)$*,* $Z_2 \sim \mathcal{N}(0, 2)$ *but they are not independent. Since conditioned on* $Z_1 = z$*, we have that* $(Z_2|Z_1 = z) = X_2 + z \sim \mathcal{N}(z, 1)$*.*

Let $\sigma_{11} = \text{Cov}(Z_1, Z_1), \sigma_{12} = \text{Cov}(Z_1, Z_2), \sigma_{22} = \text{Cov}(Z_2, Z_2)$. *With little calculation, we can get that the covariance matrix is*

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

For $Z_1|Z_2 = z$, *we show it by the equality*

$$\begin{bmatrix} a & b \\ b & c \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 \\ -\frac{b}{c} & 1 \end{bmatrix} \begin{bmatrix} \left(a - \frac{b^2}{c}\right)^{-1} & 0 \\ 0 & \frac{1}{c} \end{bmatrix} \begin{bmatrix} 1 & -\frac{b}{c} \\ 0 & 1 \end{bmatrix}$$

We can now plug it into our density function of $f_{Z_1,Z_2}(z_1, z_2)$, *the trick here is that as long as we know that we are dealing with Gaussians, we don't have to care too much about the constants because they are going to eventually work out themselves.*

$$\begin{aligned} f_{Z_1,Z_2}(z_1, z_2) &\propto \exp\left(-\frac{1}{2}\begin{bmatrix} z_1 & z_2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\frac{\sigma_{12}}{\sigma_{22}} & 1 \end{bmatrix} \begin{bmatrix} \left(\sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}\right)^{-1} & 0 \\ 0 & \frac{1}{\sigma_{22}} \end{bmatrix} \begin{bmatrix} 1 & -\frac{\sigma_{12}}{\sigma_{22}} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}\right) \\ &\propto \exp\left(-\frac{1}{2}\begin{bmatrix} z_1 - \frac{\sigma_{12}}{\sigma_{22}} z_2 & z_2 \end{bmatrix} \begin{bmatrix} \left(\sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}\right)^{-1} & 0 \\ 0 & \frac{1}{\sigma_{22}} \end{bmatrix} \begin{bmatrix} z_1 - \frac{\sigma_{12}}{\sigma_{22}} z_2 \\ z_2 \end{bmatrix}\right) \\ &\propto \exp\left(-\frac{1}{2}\frac{(z_1 - \frac{\sigma_{12}}{\sigma_{22}} z_2)^2}{\sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}}\right) \exp\left(-\frac{1}{2\sigma_{22}} z_2^2\right) \end{aligned}$$

Marginalizing out Z_1, we have that

$$f_{Z_2} \propto \exp\left(-\frac{1}{2\sigma_{22}} z_2^2\right)$$

Then conditional on $Z_2 = z_2$, we can see that $Z_1|Z_2 = z_2 \sim \mathcal{N}(\frac{\sigma_{12}}{\sigma_{22}} z_2, \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}) = \mathcal{N}(\frac{1}{2} z_2, \frac{1}{2})$